# Machine Learning Discrimination Discovery and Mitigation

# Background:

**What** is machine learning (ML) bias?

- trained model systematically generates predictions that favor advantaged group over a disadvantaged group
  - in relation to some set of attributes (e.g. race, ethnicity, gender)

**Why** does bias matter?

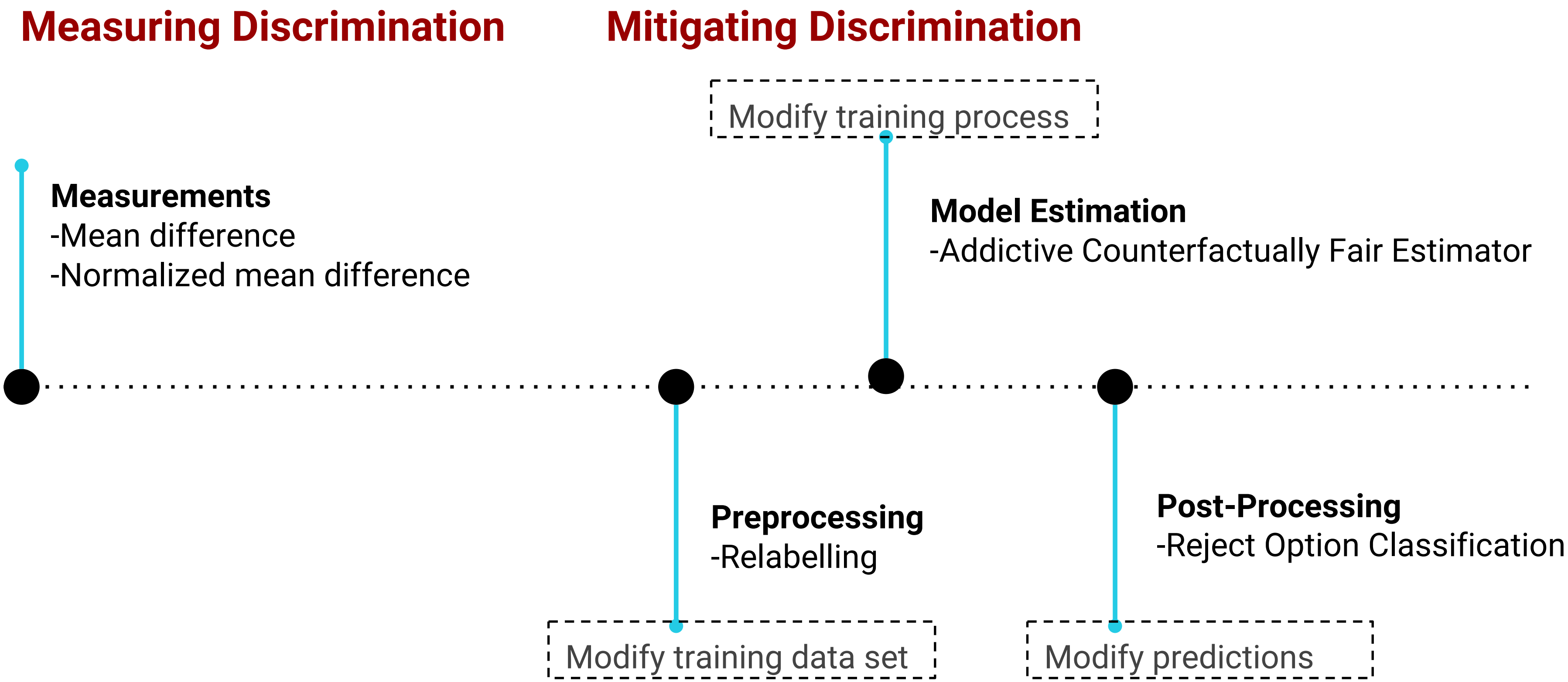
- ML is used in socially sensitive decision processes
  - hiring, loan-approval, parole-granting, etc.
- runs the risk of perpetuating socioeconomic disparities.

# Goal:

**How** do we measure and reduce bias?

- ThemisML package
  - measures and reduces potential discrimination (PD) in ML systems.
- Assumption: Protected Class: 1 for disadvantaged group and 0 for advantaged.

# Pipeline of ThemisML:



# Home Mortgage Dataset:

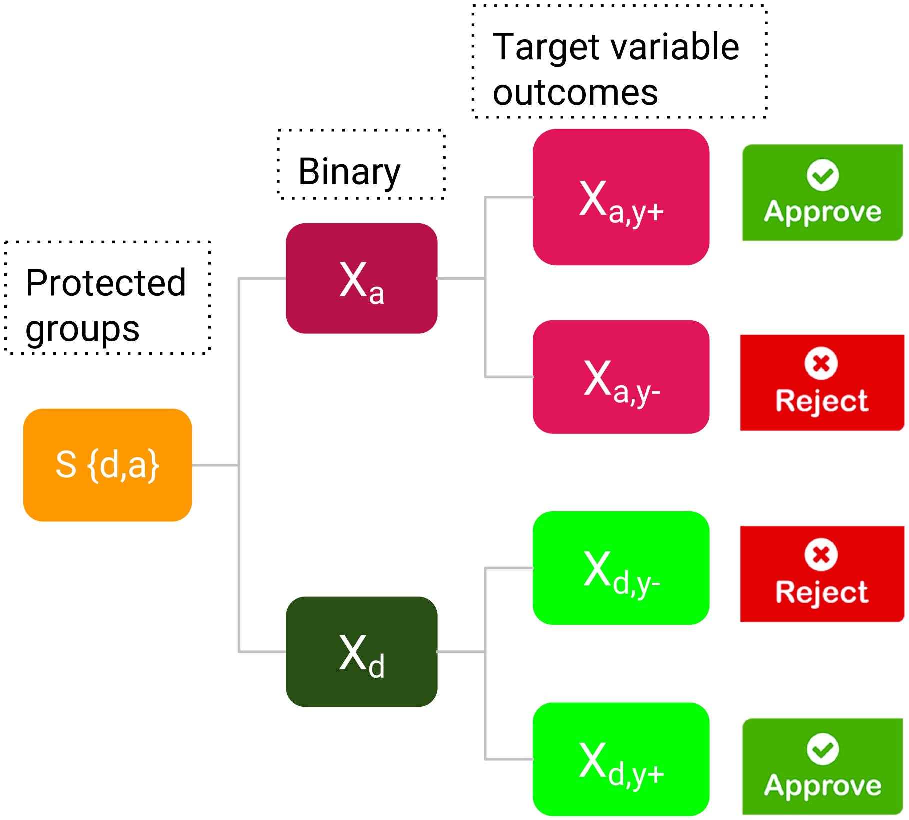
**Records of loan applications include:**

- **Loan information**:
  - Year, Agency, Loan Type, Amount, Decision
- **Applicant information**:
  - Gender, Race, Ethnicity, Occupancy, Income level
- **Loan validation**:
  - Result, Reason for denial

**Define dependent variable Y from 'Action'**

| Accept? | Loan Action |
|---|---|
| 1 | 1 -- Loan originated |
| 1 | 2 -- Application approved but not accepted |
| 0 | 3 -- Application denied |
| N/A | ~~4 --Application withdrawn by applicant~~ |
| N/A | ~~5 -- File closed for incompleteness~~ |
| 1 | 6 -- Loan purchased by the institution |
| N/A | ~~7 -- Pre Approval request denied by financial institution~~ |
| N/A | ~~8 -- Pre Approval request approved but not accepted~~ |

# Feature transformation and Metric:




-Trade-off: **Transparency** vs **Utility**

**Mean difference:**

Compute the mean difference in y with respect to protected class S

**P(y+ |Xa) - P(y+ | Xd)**

**AUC**

# Measuring Discrimination by Features:

These mean differences (MD) and confidence interval (CI) bounds suggest that on average:

- **Men** can get a loan at a *4.6% higher rate* than **women**, with a *lower bound of 4.41%* and *upper bound of 4.78%*.
- **Non-Hispanic** people get a loan at a *5.07% higher rate* than **Hispanic** people, with a *lower bound of 4.85%* and *upper bound of 5.28%*.
- **Non-Black** people get a loan at a *8.86% higher rate* than **Black** people with a *lower bound of 8.97%* and *upper bound of 25.61%*.
- **Non-American Indian/non-Alaska Native** get a loan at *15% higher rate* than American Indian/Alaska Native with a *lower bound of 14.65%* and *upper bound of 15.36%*.

| Type | Protected Class | MD (%) | MD 95% CI | Non-MD (%) | Non-MD 95% CI |
|---|---|---|---|---|---|
| Sex | Female | 4.60 | (4.41,4.78) | 4.69 | (4.50, 4.87) |
| Ethnicity | Hispanic | 5.07 | (4.85, 5.28) | 5.34 | (5.13, 5.55) |
| Race | Black | 8.86 | (8.03, 9.69) | 10.67 | (9.84, 11.50) |
| Race | American Indian/ Alaska Native | 15.00 | (14.65, 15.36) | 17.45 | (17.09, 17.80) |

# Methods

## Baseline

**Train model on all available variables**

**-Mirror the Potential Discrimination pattern in the true target variable.**

**1.** Specify model hyperparameter for training models.
2. Partition training data into 10 validation folds (VF).
3. For each VF, train model on rest of VFs.
4. Evaluate performance of model by VF.
5. Choose model with best average performance.

## RPA

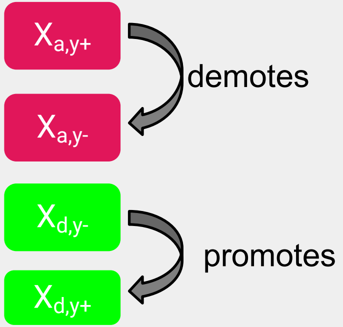
**Train model on inputs without protected attributes. I.e.** ***Naive Fairness-Aware Model***

| Protected Class |
| --- |
| Female |
| Hispanic |
| Black |
| American Indian/ Alaska Native |

## RTV pre-possessing

**Train model using the Relabelling fairness-aware method. (Reweighting; Sampling )**

Generate Ranker




Then the proportion of y+ are equal on both Xa and Xd

## ROC post-possessing

**Train a model using the Reject-option classification method. (work on: Type-agnostic predictions)**

**Training initial classifier D**
Generating predicted probabilities on test set

**Computing proximity of each prediction**
Get the decision boundary learned by D

**Assign Xd as Y+ ; Xa as Y-**
within the boundary (0.5~1)

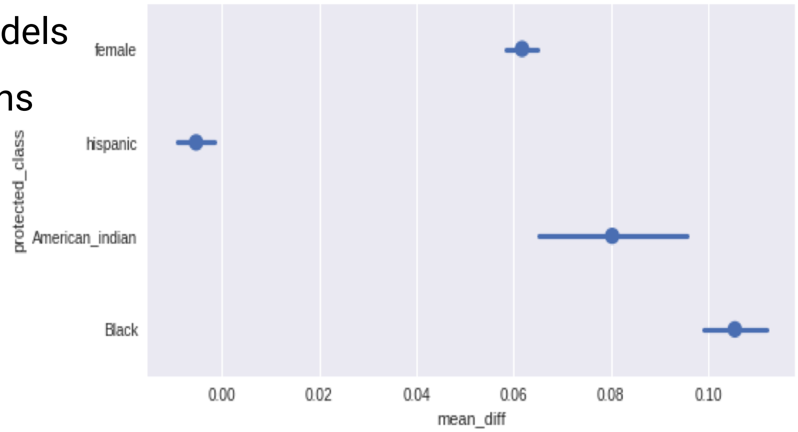
# Results from Baseline

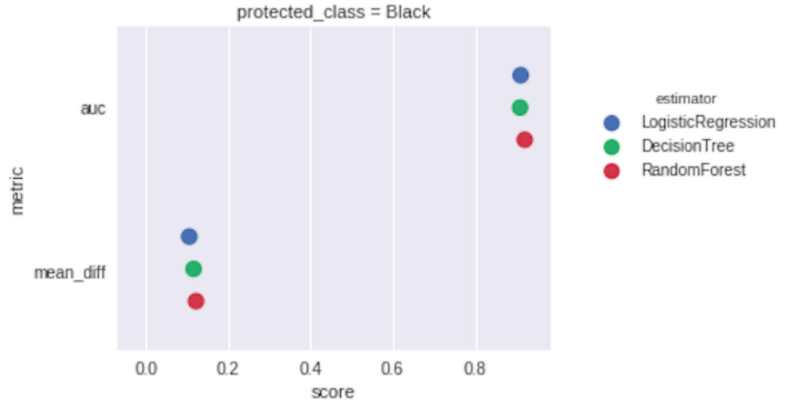
-Train Logistic Regression; Decision Tree; Random forest models

-Using 10-fold-cross validation generates train/test predictions

-AUC and mean-difference metric for each condition model

| protected_class | estimator | fold_type | auc | mean_diff | norm_mean_diff |
|---|---|---|---|---|---|
| american_indian | DecisionTree | test | 0.894434 | 0.094912 | 0.118397 |
| | LogisticRegression | test | 0.902980 | 0.099237 | 0.166110 |
| | RandomForest | test | 0.914892 | 0.113979 | 0.168575 |
| black | DecisionTree | test | 0.894028 | 0.156533 | 0.225966 |
| | LogisticRegression | test | 0.902931 | 0.162507 | 0.261264 |
| | RandomForest | test | 0.915165 | 0.171733 | 0.257836 |
| female | DecisionTree | test | 0.895168 | 0.046437 | 0.051395 |
| | LogisticRegression | test | 0.902973 | 0.049036 | 0.059731 |
| | RandomForest | test | 0.915082 | 0.051051 | 0.058372 |
| hispanic | DecisionTree | test | 0.895801 | 0.009303 | 0.011770 |
| | LogisticRegression | test | 0.902929 | 0.002552 | 0.004008 |
| | RandomForest | test | 0.915168 | 0.013858 | 0.015577 |



# Best Utility and Best Fairness

| | Female | | Black | | American Indian | | Hispanic | |
|---|---|---|---|---|---|---|---|---|
| | Mean-diff | AUC | Mean-diff | AUC | Mean-diff | AUC | Mean-diff | AUC |
| **Baseline** | **0.046** (DT) | **0.915** (RF) | **0.157** (DT) | **0.915** (RF) | **0.095** (DT) | **0.915** (RF) | **0.003** (LR) | **0.915** (RF) |
| **RPA (Naive Fairness)** | **0.046** (LR) | **0.915** (RF) | **0.156** (DT) | **0.915** (RF) | **0.085** (LR) | **0.915** ((RF) | **-0.002** (LR) | **0.915** (RF) |
| **RTV (Relabelling)** | **0.026** (DT) | **0.912** (RF) | **0.132** (DT) | **0.911** (RF) | **0.086** (LR) | **0.915** (RF) | **-0.001** (DT) | **0.914** (RF) |
| **Reject Option Classification** | **0.034** (DT) | **0.925** (RF) | **0.141** (DT) | **0.925** (RF) | **0.046** (DT) | **0.925** (RF) | **0.000** (DT) | **0.925** (RF) |

(DT: Decision Tree, RF: Random Forest, LR: Logistic Regression)

# Bias amount/Tradeoff

## Model Comparison (diff-of-diff)

| | RTV (Relabelling) | | ROC (Reject Option Classification) | |
|---|---|---|---|---|
| | Bias | AUC | Bias | AUC |
| Female | **2.0** ↓ | 0.3 ↑ | **1.5** ↓ | 1 ↑ |
| Black | **2.5** ↓ | 0.4 ↑ | **1.5** ↓ | 1 ↑ |
| A.Indian | **1.1** ↓ | - | **5.0** ↓ | 1 ↑ |
| Hispanic | **0.4(-)** ↓ | - | **0.3** ↓ | 1 ↑ |

(Unit: Percentage Point)

## The Fairness-Utility Tradeoff